基于分类权与质心驱动的无监督学习算法

刘开第 [1] 刘 昕 [2] 赵 奇 [1] 周少玲 [1]

摘 要 为了充分挖掘隐藏在样本向量中的空间信息和知识信息: 用聚类点代替类均值, 把提取指标对聚类所做贡献的量化值定义为指标分类权; 用分类权定义样本点与聚类点的加权距离, 使之作为样本与类之间的相似性度量更具合理性, 即将加权距离转化为样本隶属度. 为了消除序贯算法产生的随机性, 用样本的 K 类隶属度作为点质量的样本质点组的质心, 修正当前的 K 类聚类点, 由此建立基于分类权和质心驱动的搜索聚类点的迭代算法. IRIS 数据检验结果表明, 新算法的聚类效果与稳定性都优于已有的无监督学习方法.

关键词 无监督数据, 聚类点聚类, 分类权, 加权距离, 质心
中图分类号 TP182

An Unsupervised Learning Algorithm Based on Classification Weight and Mass Center Driving

LIU Kai-Di[1] LIU Xin[2] ZHAO Qi[1] ZHOU Shao-Ling[1]

Abstract In order to find space information and knowledge in sample points: when clustering point replaces class-mean clustering, the quantized value that describes index contribution to clustering is abstracted, then index classification weight is defined. By using classification weight, weighted distance between sample point and clustering point is defined. As similarity measurement between sample point and class, this distance is more reasonable. Transform weighted distance into sample membership. In order to avoid randomicity caused by sequential algorithm, the mass center of the sample point set is utilized to modify the present clustering points of K classes and the sample points use K memberships as their masses. From this, an iterative algorithm based on classification weight and mass center driving for searching clustering points is proposed. IRIS is used to verify this algorithm and the result shows that clustering effect and stability are superior to the existing unsupervised learning algorithms.

Key words Unsupervised data, clustering method based on clustering point, classification weight, weighted distance, mass center

支持向量机 (Support vector machine, SVM)[1] 方法的出现, 使得模型选择、过学习、非线性、维数灾难、局部极小点等困扰机器学习的问题都在很大程度上得到解决, 很多传统的机器学习方法都可看作是支持向量机方法的一种实现. 所以, 尽管支持向量机在理论与方法上还存在很多需要深入研究的东西, 但不容否认的是, 支持向量机的出现大大提升了有监督模式识别的能力.

但是, 无监督模式识别远没有这样乐观. 因为分类信息太少, 有效的学习方法原本就不多, 并且十多年来这种现状少有改观. 虽说有自组织映射 (Self-organizing map, SOM)[2]、模糊 SOM、K-means 和模糊 K-means 等学习方法[3], 但是, 反映新思路、新方法的更有效的学习算法尚属鲜见. 这种算法滞后和无监督数据急剧膨胀的现状急需改变, 所以关注无监督学习算法研究十分迫切.

不同指标在分类中起的作用不同, 有大有小. 若去掉作用非常小的指标, 那么降维后对分类不会有多少影响. 但是, 若去掉一项作用很大的指标, 将严重影响分类结果. SOM 的特点是把高维空间中的样本点映射到二维平面上, 借用二维平面的可视性便于对样本点分类. 这种盲目地、大幅度地降维势必造成严重的分类信息失真. 从表 1 (见第 529 页) 中各学习方法的学习效果看, SOM 方法效果最差, 其原因就在于此. 所以, 要想得到好的学习效果, 盲目降维和大幅度降维的学习方法都是不可取的.

K-means 方法在不降维条件下得到的是在误差平方和最小意义上的最优聚类. 不足在于: 1) 用类

收稿日期 2008-01-21 收修改稿日期 2008-10-23
Received January 21, 2008; in revised form October 23, 2008
国家自然科学基金 (60474019), 河北省自然科学基金 (F2005000482) 资助
Supported by National Natural Science Foundation of China (60474019) and Natural Science Foundation of Hebei Province (F2005000482)
1. 河北工程大学不确定性数学研究所 邯郸 056038 2. 中国矿业大学 (北京) 化环学院 北京 100083
1. Institution of Uncertainty Mathematics, Hebei University of Engineering, Handan 056038 2. School of Chemical and Environmental Engineering, China University of Mining and Technology (Beijing), Beijing 100083
DOI: 10.3724/SP.J.1004.2009.00526

均值代表点代表类从分类角度讲并非最优; 2) 用样本点到类均值点的欧氏距离作为样本点与类之间的相似性度量, 并按最小距离准则对样本点归类, 从分类角度讲并不合理; 3) 逐点修正类均值的序贯算法的随机性影响解的稳定性. 这些不足, 使得基于 K-means 聚类的学习效果也不够理想, 但是, 这种不足也为设计新的学习方法保留了一定的生存空间.

从 SOM 和 K-means 算法得到这样的启示: 若想提高无监督学习的学习效果, 那么, 在尽可能不降维条件下, 必须充分挖掘隐藏在样本向量中对样本分类有利的启发性知识用于指导样本分类; 尽可能用批处理算法替代序贯算法, 以便消除随机性, 提高解的稳定性, 同时要充分利用样本集在空间中分布的整体特性. 按照上述启示, 我们设计了一种基于分类权与质心驱动的无监督学习算法.

为了能在 d 维空间中定义两点间的某种距离, 须将 d 维实测参数空间转化为 d 维标称化指标空间. 如令

$$y_{ij}=\frac{x_{ij}-\min\limits_{i}\{x_{ij}\}}{\max\limits_{i}\{x_{ij}\}-\min\limits_{i}\{x_{ij}\}},i=1,\cdots,N,j=1,\cdots,d \tag{1}$$

其中, x_{ij} 是样本 x_i 关于参数 j 的实测值, 变换后的 y_{ij} 位于单位超立方体上.

上述无量纲化过程, 势必造成一定程度的分类信息失真, 但是, 对于无监督学习这是不得已的.

1　无监督学习中的分类数与学习思路

无监督学习的分类数在理论上没有规范的确定方法. 但是, 可利用 SOM 在二维平面上的可视性, 或 K-means 聚类中 (误差−聚类数) 曲线上的拐点, 或必要的领域知识确定无监督样本的大致分类数; 并且当确定了大致分类数后还可通过试分类的方法选择更合适的分类数. 所以, 不妨假定分类数 P 是已知的.

与监督学习不同, 无监督学习没有已知分类的样本供学习使用. 如果说监督学习是设法揭示同类样本在空间中的分布规律, 并把这种规律作为识别待识样本类别的依据, 那么, 无监督学习则是不管 N 个样本点在空间中实际上遵循怎样的规律自动分为 P 个类, 一律按 “分量相对接近的样本点归为一类” 的原则, 把 N 个样本点划分为 P 个类. 为了搜索到 N 个样本点在空间中形成的 P 个相对集中的区域, 有下述学习思路:

1) 按照 N 个样本点各维分量相对接近程度, 将样本集划分为 P 个初始分类, 每一类中样本的各维分量是相对接近的.

2) 为了提取隐藏在样本向量中有利于样本分类的数据信息, 用代表点代表类时提取各指标对分类所做贡献的量化值, 以此为启发性知识定义样本点的各类隶属度.

3) 为了用批处理法替代序贯算法以消除随机性, 以样本点的 K 类隶属度作为点质量的 N 个样本点构成的质点组的质心, 修正当前 K 类代表点, 则各样本点在新的类代表点下对应新的隶属度; 而新的隶属度生成新的质心. 由此可建立基于质心驱动的搜索类代表点的迭代算法.

4) 由迭代算法确定的 P 个类代表点, 给出 N 个样本点的 P 个聚类.

2　初始分类与指标分类权

无监督学习是按照分量的接近程度对样本点分类, 所以, 要求初始分类的每一类的样本点其各维分量尽可能相对接近. 按照这种要求给出下述初始分类方法.

设 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}\,(i=1,\cdots,N)$ 是 d 维标称化指标空间中的第 i 个样本, 令

$$SUM(i)=\sum_{j=1}^{d}x_{ij} \tag{2}$$

$$MA=\max_{i}SUM(i) \tag{3}$$

$$MI=\min_{i}SUM(i) \tag{4}$$

$$J=\frac{(P-1)\cdot[SUM(i)-MI]}{MA-MI} \tag{5}$$

若 K 是与 $J+1$ 最接近的整数, 则将样本 x_i 归入第 K 类. 这样, N 个样本刚好被划分为 P 个不同类. 用 $c_k(0)$ 表示第 $k\,(k=1,\cdots,p)$ 个初始分类, 内含 $n_k(0)$ 个样本点, 其均值记为 $m_k(0)$.

以 $m_k(0)$ 为初始代表点代表 $c_k(k=1,\cdots,p)$ 理想类. 这样, 当用 P 个代表点去代表 P 个类时, 则这 P 个类能否被区分开和在怎样程度上被区分开取决于这 P 个点能否被区分开为 P 个不同点和在什么样的程度上被区分开为 P 个不同点, 而后者由各点的 d 个分量完全决定. 所以, 不同指标在分类中起的作用不同, 或者说对分类所做的贡献不同. 因此, 要想对样本正确分类, 就不能不考虑各项指标在分类中的作用互不相同这一基本事实; 必须提取各维指标对分类所做贡献的量化值用于指导样本分类.

为此, 令

$$m_k(0)=[m_{k1}(0),m_{k2}(0),\cdots,m_{kd}(0)],\ k=1,\cdots,p \tag{6}$$

$$\overline{\boldsymbol{m}}(0)=\frac{1}{p}\sum_{k=1}^{p}m_k(0)=[\overline{\boldsymbol{m}}_1(0),\overline{\boldsymbol{m}}_2(0),\cdots,\overline{\boldsymbol{m}}_d(0)] \tag{7}$$

$$\boldsymbol{\delta}_j^2(0)=\frac{1}{p}\sum_{k=1}^{p}[m_{kj}(0)-\overline{\boldsymbol{m}}_j(0)]^2,\ j=1,\cdots,d \tag{8}$$

$$\boldsymbol{\alpha}_j(0)=\frac{\boldsymbol{\delta}_j^2(0)}{\sum\limits_{t=1}^{d}\boldsymbol{\delta}_t^2(0)} \tag{9}$$

显然 $\boldsymbol{\alpha}_j(0)$ 满足:

$$0\le\boldsymbol{\alpha}_j(0)\le 1,\quad \sum_{j=1}^{d}\boldsymbol{\alpha}_j(0)=1 \tag{10}$$

称 $\boldsymbol{\alpha}_j(0),\ j=1,\cdots,d$ 为 j 指标在以 $m_1(0),m_2(0),\cdots,m_p(0)$ 为类代表点条件下的分类权.

如果分类权 $\boldsymbol{\alpha}_j(0)=0$, 则 P 个代表点 $m_1(0),m_2(0),\cdots,m_p(0)$ 的 j 维分量都相同. 这说明 j 指标对于把 P 个代表点区分开不起作用, 即 j 指标对于把 P 个类区分开不起作用, 删除 j 指标也不会影响分类. 所以, 当分类权 $\boldsymbol{\alpha}_j(0)=0$ 时, 表明 j 指标是对分类不起作用的冗余指标.

3 加权距离与样本隶属度

如果用点 $m_k(0),\ k=1,\cdots,p$ 作为类代表点代表 c_k 类, 用样本点 y_i 与代表点 $m_k(0)$ 的某种距离 $d[y_i,m_k(0)]$ 作为样本点 y_i 与 c_k 类之间的相似性度量, 并且, 按 "最小距离" 准则对样本分类. 那么, 由于 $d[y_i,m_k(0)]$ 中包含着 y_i 的分类信息, 可知 $d[y_i,m_k(0)]$ 不是欧氏距离. 比如, 当分类权 $\boldsymbol{\alpha}_j(0)=0$ 时, 因为 j 指标对分类不起作用, 删除 j 指标也不会影响分类. 所以此种情况下, j 分量就不应出现在计算 $d[y_i,m_k(0)]$ 的相应公式中. 可见, $d[y_i,m_k(0)]$ 是一种以区分权 $\boldsymbol{\alpha}_j(0)$ 为权的加权距离:

$$[d(y_i,m_k(0))]^2=\sum_{j=1}^{n}\boldsymbol{\alpha}_j(0)\cdot[y_{ij}-m_{kj}(0)]^2 \tag{11}$$

显然, $d[y_i,m_k(0)]$ 越小, 则样本 y_i 隶属于 c_k 类的可能性就越大. 如果用 $\mu_k^{(0)}(y_i)$ 表示 y_i 属于 c_k 类的初始隶属度, 则 $\mu_k^{(0)}(y_i)$ 可表示为

$$\mu_k^{(0)}(y_i)=\frac{\dfrac{1}{\delta+d[y_i,m_k(0)]}}{\sum\limits_{t=1}^{p}\dfrac{1}{\delta+d[y_i,m_t(0)]}} \tag{12}$$

其中, δ 为控制常数.

这样, 如果利用 P 个点 $m_1(0),m_2(0),\cdots,m_p(0)$ 作为类代表点分别代表 P 个类 $c_1,\cdots,c_p$, 那么, 任意可能的样本点 $y_i,\ i=1,\cdots,N$ 属于 c_k 类的隶属度可由式 (12) 确定.

注意到我们的目的是想知道: N 个样本点在空间中 "相对集中" 的 P 个区域, 而用逐个输入样本点调整类均值的序贯算法, 因无法克服的随机性影响解的稳定性, 且不利于充分利用样本点集在空间中分布的整体特性. 那么, 怎样的启发性知识能指示搜索第 K, $K=1,\cdots,p$ 个区域的搜索方向呢?

注意到 K 类样本点相对集中的区域一定是以 K 类隶属度为点质量的样本点构成的质点组的质心所在的区域, 而以 K 类隶属度作为点质量的 N 个样本点构成的质点组的质心是可以确定的. 这样, 可用质心去修正当前的 K, $K=1,\cdots,p$ 类代表点. 由此建立基于质心驱动的搜索类代表点的迭代算法, 进而用批处理算法替代序贯算法, 消除随机性.

4 搜索类代表点的迭代算法

4.1 迭代算法步骤

步骤 1. 设 c_k 类的初始类代表点为 $m_k(0),\ k=1,\cdots,p$, 样本 y_i 属于 c_k 类的初始隶属度为 $\mu_k(y_i,0),\ k=1,\cdots,p,\ i=1,\cdots,N$;

步骤 2. 迭代按节拍 t 进行, 置 $t=1$. 最大迭代次数为 $t_{\max}$, 终止常数 $\varepsilon>0$;

步骤 3. 计算当前类代表 $m_1(t-1),m_2(t-1),\cdots,m_p(t-1)$ 条件下, j 指标的分类权 $\boldsymbol{\alpha}_j(t-1),\ j=1,\cdots,d$;

步骤 4. 计算样本 y_i 到代表点 $m_k(t-1)$ 的加权距离 $d[y_i,m_k(t-1)],\ k=1,\cdots,p,i=1,\cdots,N$;

步骤 5. 用加权距离计算样本 y_i 属于 c_k 类的隶属度 $\mu_k(y_i,t-1),\ k=1,\cdots,p,i=1,\cdots,N$;

步骤 6. 将 $\mu_k(y_i,t-1)$ 作为点质量赋予点 y_i, 计算由 N 个样本点构成的质点组的 K 类质心:

$$O_k(t-1)=\frac{\sum\limits_{i=1}^{N}\mu_k(y_i,t-1)\cdot y_i}{\sum\limits_{i=1}^{N}\mu_k(y_i,t-1)},\quad k=1,\cdots,p$$

步骤 7. 按照下述公式修正当前类代表点:

$$m_k(t)=m_k(t-1)+\frac{w(t)}{t}[O_k(t-1)-m_k(t-1)],\qquad k=1,\cdots,p$$

其中 $w(t)$ 是 t 的单减函数, 比如取

$$w(t)=\begin{cases}0.5, & t=1\\ \dfrac{0.5}{t^2\sqrt{t}}, & 2\le t\le t_{\max}\end{cases}$$

步骤 8. 计算并比较

$$\sum_{k=1}^{p}\sum_{j=1}^{d}[m_{kj}(t)-m_{kj}(t-1)]^2<\varepsilon?$$

若否, 继续; 若是, 转步骤 10;

步骤 9. $t<t_{\max}$? 若是, 令 $t=t+1$, 转步骤 3; 否则, 转步骤 10;

步骤 10. 停止, 输出:

1) 当前类代表点 $m_1(t)\sim m_p(t)$;

2) 当前类代表点条件下的分类权 $\boldsymbol{\alpha}_1(t)\sim\boldsymbol{\alpha}_d(t)$;

3) 计算样本 y_i 到各类代表点的加权距离, 并按 "最小加权距离" 准则将 N 个样本归类;

4) 计算各样本点关于各类的隶属度.

4.2　算法讨论

1) 随着节拍 t 增大, $m_k(t-1)$ 移动的距离逐渐减小且最终趋于 0, 所以迭代算法收敛.

2) 按样本的各维分量接近程度划分的理想类 c_k 与 N 个样本生成的 "真实第 K 类" 并非完全一致. 由 "质心驱动" 获得的代表点生成的类只是理想 c_k 类的一种近似, 更是 "真实第 K 类" 的近似. 所以, 由迭代算法得到的聚类是满意聚类, 无法找到实际上的最佳聚类.

3) 无监督聚类的聚类规则是把分量相对接近的样本归为一类. 所以, 基于质心驱动的搜索类代表点的学习算法要求: 初始分类中每一类样本的各维分量应尽可能是相对接近的. 这是获得较好学习效果的必要条件.

4) 学习的目的是搜索 N 个样本点在空间中按各维分量 "相对接近" 程度划分的 P 个聚类区域, 这个区域实际上是无法确切知道的, 而用于代表 K 类区域的代表点也是无法确切知道的. 所以增加迭代次数主要是为了得到稳定的解, 并不意味着迭代次数越多, 解的质量一定越好.

5　算法有效性检验

任何无监督学习算法是否有效必须经过有效性检验才知道. IRIS 数据[4] 是国际上公认的检验无监督聚类效果的检验数据. IRIS 数据分为三类, 每类 50 个样本, 每个样本都是关于花瓣测量值的 4 维数据.

5.1　检验方法

1) 按式 (1) 将 IRIS 数据转化为 4 维标称化数据;

2) 按式 (2)～(5) 将 IRIS 数据划分为三个初始分类, 用类均值 $m_1(0), m_2(0), m_3(0)$ 作为初始代表点, 分别代表 c_1, c_2, c_3 理想类;

3) 按迭代算法搜索类代表点. 输出类代表点并计算错分样本数.

不同学习算法关于 IRIS 数据的检验结果如表 1.

表 1　检验结果比较

Table 1　Comparison of test results

学习方法	错分样本数	稳定性	备注
SOM	不少于 27 个	结果不确定	见文献 [5]
K-means	不少于 16 个	结果不确定	见文献 [5]
Nearal Gas	不少于 11 个	结果不确定	见文献 [5]
Ng-jordan	不少于 16 个	结果不确定	见文献 [5]
文献 [5]	不少于 7 个	结果不确定	
文献 [6]	不少于 7 个	结果不确定	
本文算法	5 个	结果确定	

由表 1 看出, 本文算法的学习效果和稳定性明显优于其他无监督学习方法. 而且算法简单、可重复、收敛速度快, 并具有实时性强的特点.

5.2　检验过程

1) 按式 (2)～(5) 将 IRIS 数据划分为三个初始分类 Ⅰ、Ⅱ、Ⅲ. Ⅰ类含样本 41 个; Ⅱ类含样本 88 个; Ⅲ 类含样本 21 个.

错分样本数为 44 个, 其中 12 个由 Ⅰ 错分到 Ⅱ; 3 个由 Ⅱ 错分到 Ⅰ; 29 个由 Ⅲ 错分到 Ⅱ.

2) 三个初始分类的类均值为

$$m_1(0)=(0.1660,\ 0.4959,\ 0.0992,\ 0.0703)$$
$$m_2(0)=(0.4684,\ 0.3996,\ 0.5487,\ 0.5374)$$
$$m_3(0)=(0.7751,\ 0.4940,\ 0.8467,\ 0.8750)$$

以类均值为初始代表点, 按最小加权距离准则对 150 个样本重新分类, 则错分样本为 14 个.

把初始分类的类均值作为初始类代表点. 当以加权距离作为样本点与类之间的相似性度量时, 在 "最小距离" 识别准则下, 错分样本由 44 个降为 14

个. 说明 “加权距离” 作为样本点与类之间的相似性度量对于改善分类效果具有实质上的重要性.

3) 用迭代法求类代表点

当确定了初始类代表点后, 对迭代结果影响最大的是参数 $w(t)$, 对于不同的 $w(t)$, 迭代结果以及趋于稳定所需的迭代次数不同. 比如, 取 $\delta = 0.00005$ 时

a) 取

$$w(t) = \begin{cases} 0.5, & t = 1 \\ \dfrac{0.5}{t\sqrt{t}}, & 2 \le t \le t_{\max} \end{cases}$$

则迭代四次以后, 错分数稳定在 7 个.

b) 取

$$w(t) = \begin{cases} 0.5, & t = 1 \\ \dfrac{0.5}{t^2\sqrt{t}}, & 2 \le t \le t_{\max} \end{cases}$$

则第三次迭代错分数为 5 个, 一直迭代 200 次, 错分数都稳定在 5 个. 迭代三次后的代表点为

$$m_1(3) = (0.2211,\ 0.5121,\ 0.1599,\ 0.1386)$$

$$m_2(3) = (0.4680,\ 0.3828,\ 0.5556,\ 0.5393)$$

$$m_3(3) = (0.6724,\ 0.4534,\ 0.7543,\ 0.7749)$$

分类权为

$$\boldsymbol{\alpha}(t) = (\boldsymbol{\alpha}_1(3),\ \boldsymbol{\alpha}_2(3),\ \boldsymbol{\alpha}_3(3),\ \boldsymbol{\alpha}_4(3)) = (0.2040,\ 0.0168,\ 0.3659,\ 0.4136) \quad (13)$$

误差为

$$\sum_{k=1}^{3}\sum_{j=1}^{4}[m_{kj}(3) - m_{kj}(2)]^2 = 2.3064 \times 10^{-5}$$

迭代 200 次后代表点为

$$m_1(200) = (0.2220,\ 0.5123,\ 0.1610,\ 0.1397)$$
$$m_2(200) = (0.4679,\ 0.3827,\ 0.5553,\ 0.5390)$$
$$m_3(200) = (0.6705,\ 0.4528,\ 0.7526,\ 0.7730)$$

错分样本为 5 个. 误差为

$$\sum_{k=1}^{3}\sum_{j=1}^{4}[m_{kj}(200) - m_{kj}(199)]^2 = 3.7759 \times 10^{-18}$$

所以, 适当选择步长参数 $w(t)$ 对改善聚类效果十分重要.

5.3 初始分类对学习效果的影响

基于质心驱动的学习算法, 要求初始分类中每一类样本点的各维分量应尽可能是 “相对接近” 的, 这样才能保证有较好的学习效果. 如果用随机法分类, 因为不能保证同类中样本的各维分量是 “相对接近” 的, 所以, 基于质心驱动的迭代算法不会有好的学习效果, IRIS 数据的检验结果证实了这一点. 但是, 只要初始分类满足同类样本点的各维分量 “相对接近” 的条件, 那么, 基于质心驱动的学习算法都会有较好的学习效果. 比如, 改用 “密度” 法确定初始分类:

1) 称样本点 y_i $(i = 1, \cdots, 150)$ 的 δ 邻域内包含的样本数为 y_i 的密度;

2) 具有最大密度的样本点作为第一个初始类代表点, 记为 $m_1(0)$;

3) 在点 $m_1(0)$ 的 $\eta(\eta \ge \delta)$ 邻域以外选择最大密度点作为第二个初始类代表点, 记为 $m_2(0)$;

4) 点 $m_1(0)$ 和 $m_2(0)$ 的 η 邻域外的最大密度点作为第三个初始类代表点, 记为 $m_3(0)$;

5) 以点 $m_1(0)$, $m_2(0)$, $m_3(0)$ 作为类代表点. 当取 $\delta = 0.14, \eta = 0.15$ 时得到的初始类代表点, 在加权距离下的错分样本为 12 个;

6) 迭代算法后, 错分样本个数稳定在 6 个.

6 结论

1) IRIS 数据检验结果表明, 用样本点与代表点间的加权距离作为样本点与类之间的相似性度量时, 分类效果明显优于欧氏距离. 说明提取指标对分类所做贡献的量化值用于指导样本分类对改善学习效果具有实质上的重要性.

2) 基于分类权与质心驱动的学习方法是一种批处理法, 因消除了样本输入阶段产生的随机性, 所以对 IRIS 数据的聚类结果稳定并且聚类效果明显优于其他学习方法, 说明这是一种有效的无监督学习方法.

3) 基于分类权和质心驱动的无监督学习策略是: 分量相对接近的样本归为一类. 所以, 算法适用于呈球形分布或近似呈球形分布的无监督样本数据.

4) 质心驱动算法应注意选择合适的步长 $w(t)$, 且初始类代表点选择的是否合适, 对解有较大影响.

5) 由式 (13) 可知, $\boldsymbol{\alpha}_2(3) = 0.0168$, 相对甚小, 说明 IRIS 数据的第二项指标对聚类影响很小. 去掉第二项指标对三维数据进行聚类, 所得聚类

结果基本相同. 这说明适当降维是可行的, 但不能盲目.

References

1 Vapnik V N. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995

2 Kohonen T, Oja E, Simula O, Visa A, Kangas J. Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 1996, **84**(10): 1358−1384

3 Bian Zhao-Qi, Zhang Xue-Gong. *Fuzzy Recognition*. Beijing: Tsinghua University, 2000. 236−280
(边肇祺, 张学工. 模糊识别. 北京: 清华大学出版社, 2000. 236−280)

4 Everitt B S, Landau S, Leese M. *Cluster Analysis (Third Edition)*. New York: Halsted Press, 1993

5 Camastra F, Verrian A. A novel kernel method for clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(5): 801−805

6 Newton S C, Surya P, Sunanda M. Adaptive fuzzy leader clustering complex data sets in patten recognition. *IEEE Transactions on Neural Networks*, 1992, **3**(5): 794−800

刘开第 河北工程大学不确定性数学研究所教授. 主要研究方向为不确定性信息处理. 本文通信作者.
E-mail: liukaidi@hebeu.edu.cn
(**LIU Kai-Di** Professor at the Institution of Uncertainty Mathematics, Hebei University of Engineering. His main research interest is processing method of unascertained information. Corresponding author of this paper.)

刘 昕 中国矿业大学 (北京) 化环学院讲师. 主要研究方向为信息处理.
E-mail: liuxin@hebeu.edu.cn
(**LIU Xin** Lecturer at the School of Chemical and Environmental Engineering, China University of Mining and Technology (Beijing). His main research interest is processing method of unascertained information.)

赵 奇 河北工程大学不确定性数学研究所教授. 主要研究方向为不确定性信息处理. E-mail: zhaoqi@hebeu.edu.cn
(**ZHAO Qi** Professor at the Institution of Uncertainty Mathematics, Hebei University of Engineering. His main research interest is processing method of unascertained information.)

周少玲 河北工程大学不确定性数学研究所讲师. 主要研究方向为不确定性信息处理. E-mail: zhoushaoling@sina.com
(**ZHOU Shao-Ling** Lecturer at the Institution of Uncertainty Mathematics, Hebei University of Engineering. Her main research interest is processing method of unascertained information.)